인공지능 문명 전환기의 위기 진단과 원불교의 AI윤리

김 화 연

<목 차>

Ⅰ. 문제제기: 인류는 지금 어디에 서 있는가
Ⅱ. 개인의 기회와 위기
Ⅲ. 사회의 기회와 위기
Ⅳ. 인류·지구의 기회와 위기
Ⅴ. AI 위험의 발생·증폭 조건과 전환의 갈림길
Ⅵ. 브레이크 없는 전환기, 종교가 할 수 있는 일
Ⅶ. 원불교 AI윤리 원칙
참고문헌

Ⅰ. 문제 제기: 인류는 지금 어디에 서 있는가

1. 두려움에서 시작하다

AI 전환기인 2026년, 인류는 이전과는 다른 차원의 두려움을 마주하고 있다. 그 두려움은 다음과 같은 질문으로 압축할 수 있다. “AI가 나의 일자리를 빼앗는 것은 아닌가?” “인간은 결국 쓸모없어지는 것이 아닌가?”

일자리는 기술과 함께 바뀌어 왔다. 산업혁명 이후 사람들은 농촌에서 공장으로, 공장에서 사무실로, 반복적인 사무에서 전문직과 서비스로 옮겨 갔다. 기계가 정해진 방식으로 되풀이되는 일을 떠맡을 때마다, 사람은 기계가 아직 하지 못하는 다른 일로 이동했다. 그 과정에서 생산력이 크게 늘었고, 인류는 절대적 빈곤에서 벗어나 물질적 번영을 누릴 수 있었다. 그러나 지금의 AI 발전은 인간의 고유 영역으로 여겨졌던 지적 전문 업무까지 대체한다는 점에서 이제까지의 기술 발전과는 다른 양상을 띤다.

두려움은 여기서 멈추지 않고, 인간의 통제를 벗어나려는 기술 그 자체로 향한다. 이 두려움은 막연한상상으로 치부할 수 없다. AI를 이용한 사기·허위정보·보안 사고와 오작동 피해가 이미 보고되고 있다. 그러나 위기가 감지되고 있음에도 불구하고 AI 경쟁우위 확보를 위한 기업, 국가간 패권 경쟁은 안전대비를 어렵게 만들고 있다.

2. 두려움 너머의 불안 – 실존의 위기

일자리를 잃을지 모른다는 두려움이 개인의 위기라면, AI가 인간의 통제와 감독을 벗어나 큰 사고를 일으킬 수 있다는 두려움은 인류 전체에게 해당하는 위기다. 이 위기는 생계와 안전이라는 삶의 조건을 위협하는 동시에, 인간이 자기 자신을 이해해 온 근거를 뒤흔든다.

인간은 오랫동안 무슨 일을 하고 무엇을 할 수 있는가로 자신을 소개하고 이해해 왔다. 더욱이 인간의 가치를 생산성과 유용성, 타인에게 제공하는 기능으로 측정해온 자본주의 사회에서 일이 없다는 것은 곧 쓸모없다는 판정으로 연결된다. 일은 그동안 그 사람이 누구인지를 설명해 주었고, 그 사람의 쓸모를 증명해 주었으며, 하루와 삶의 방향을 정해 주었다. 그래서 일을 잃는 순간, ‘무엇을 하는가’라는 물음은 ‘나는 누구인가’, ‘나는 무엇을 해야 하는가’, ‘무엇을 위해 살아야 하는가’라는 물음으로 전환된다. 사회 제도가 개선되어 일자리를 잃은

이들의 생계를 받쳐 주더라도 이러한 실존적 고민에 대해서는 답해 주지 못한다.

3. 기회와 위험

AI 발전은 인류에게 공멸이라는 위기를 불러오는 동시에, 인류 문명의 비약적 발전이라는 기회를 의미한다. AI기술이 누구에게 어떤 권한으로, 어떤 목적과 조건 아래 쓰이느냐에 따라 인류가 오랫동안 풀지 못한 난제를 푸는 기회가 될 수도, 감시와 차별, 공멸의 수단이 될 수도 있다. AI기술이 가져올 수 있는 가능성은 영향이 미치는 범주에 따라 개인, 사회, 지구와 인류의 세 범위로 나누어 살피고, 각 범위를 다시 물질과 정신(문화)의 두 측면으로 나누어 살펴 본다.(영육쌍전을 기준으로 적용) 물질적 측면은 노동·안전·생태 자원 등 인간의 생존과 번영에 관계된 물질적 경제적 조건 전반을 포함한다. 정신문화적 측면은 개인에서는 정신, 사회공동체에서는 정신문화적 측면을 포괄한다. 사회적 범주에서는 구성원 사이의 공동체의 신뢰와 유대, 공론장, 사회 규범 등의 문화, 인류 전체의 범주에서는 서로 다른 문화와 문명의 존중, 교류와 소통, 공존의 조건을 포괄한다.

Ⅱ. 개인의 기회와 위기

이 장은 AI가 가져올 수 있는 기회와 위기를 개인적 차원에서 물질적 조건과 정신의 영역, 두 측면으로 살펴본다.

개인 차원의 기회와 위기

		기회	위기
개인	물질·경제	반복노동과 위험한 현장 업무의 대체 업무의 효율 증진 의료 혜택의 확장: 진료 접근성 확대와 개인 맞춤형 진료	일과 소득 불안 건강·금융·안전 판단의 오류 개인정보 침해와 조작된 음성·영상을 이용한 범죄
	정신	반복 노동에서 풀려나 확보된 여유 맞춤형 배움, 지식 접근 및 분석, 창작 보조 언어와 장애의 벽을 넘는 표현과 소통	주의와 정서의 포획(맞춤형 자극) 정서적 의존과 직접적 인간 관계의 위축 사고와 판단·결정의 위임(과신, 탈숙련, 자기결정의 약화) 일과 역할을 잃은 뒤의 의미 상실 존재 가치의 위기와 도피

1. 물질·경제적 측면의 기회와 위기

기회: AI는 이미 법률 문서를 검토하고, 의료 영상을 판독하며, 코드를 작성하고 수정한다. 이제는 답을 제시하는 도구를 넘어, 과제를 여러 단계로 나누고 하위 작업을 계획·조정하며 외부 도구를 사용해 일정 시간 동안 실행하는 에이전트형 시스템으로 발전하고 있다. AI의 능력은

과업과 상황에 따라 아직 고르지 않지만, 여러 영역에서 사람이 수년 혹은 수십 년의 훈련을 통해 쌓은 역량을 따라잡거나 앞서고 있다.

이 변화는 물질적 조건의 비약적 발전을 의미한다. AI는 일상의 여러 업무를 보조하여 같은 시간에 더 많은 일을 해내도록 돕는다. 로봇과 자동화는 반복적이거나 위험한 현장 업무를 대신한다. 의료 영상 판독과 진료 보조는 전문의가 드문 지역과 계층에까지 진료의 문턱을 낮추고, 개개인에 최적화된 맞춤형 진료를 가능하게 한다.

위기: AI가 가져올 이익과 비용은 고르게 나뉘지 않는다. 개인에게 그 불균형은 먼저 일자리와 소득의 불안으로 다가오며, 그 조짐은 이미 청년층에서 나타나고 있다. 스탠퍼드 디지털경제연구소의 2026년 8월 분석에 따르면, 생성형 AI가 확산된 뒤 AI가 대신할 수 있는 업무가 많은 직종에서 22~25세 청년의 고용은 AI의 영향을 덜 받는 직종의 같은 나이대보다 약 19% 뒤처졌다. 경력의 첫 단계에서 맡던 업무부터 AI가 대신하면서, 청년이 일자리에 들어설 문이 먼저 좁아지고 있는 것으로 보인다. 부담은 성별에 따라서도 다르게 나타난다. 국제노동기구의 2025년 분석에서 AI의 영향을 가장 크게 받는 직종은 사무·행정직이었고, 고소득 국가에서 이 범주에 속한 일자리의 비율은 여성이 9.6%로 남성(3.5%)의 세 배 가까이 되었다. 이처럼 AI 전환에 따른 부담은 청년과 여성, 중간숙련 사무직, 재교육과 디지털 접근이 어려운 계층에 먼저 전가되고 있다.

위험은 일자리에만 있지 않다. 많은 이들이 건강·금융·법률 문제를 AI에게 묻는다. 대화가 자연스럽다고 해서 답이 정확하다는 보장은 없다. 잘못된 의료·금융·안전 정보는 직접적인 피해로 연결될 수 있다. 개인정보도 새로운 차원의 위험에 노출되어 있다. 맞춤형 서비스를 위해 음성·얼굴·건강·위치 등 개인정보가 과도하게 수집되거나 유출될 수 있고, AI로 만든 목소리와 영상은 가족이나 기관을 사칭한 사기와 협박을 더 그럴듯하게 만든다. 이러한 피해는 AI기술의 안전에 대한 검증과 피해 구제 제도의 구축보다 도입 속도가 앞설 때 더욱 커지며, 그 비용은 대개 정보와 대응 수단이 부족한 개인에게 전가된다.

2. 정신적 측면의 기회와 위기

기회: AI는 정신적 측면에서도 여러 가능성을 의미한다. 반복 노동에서 해방되며 확보한 시간은 배움과 자기 성찰, 돌봄과 관계를 위해

사용할 수 있다. 사람이 목적과 질문을 정하고 답을 검토한다면, AI는 자료를 찾아 정리하고 여러 관점을 비교하며 반복 연습을 돕는 맞춤형 학습 보조자가 될 수 있다. 번역과 음성 인식, 화면 해설은 언어와 장애의 벽을 낮추어 더 많은 사람이 자신을 표현하고 소통할 수 있게 한다.

위기

가. 주의와 정서의 포획: AI 기반 추천 시스템은 이용자의 선택 기록과 반응을 분석해 관심을 끌 만한 주장·콘텐츠·상품을 끊임없이 내놓는다. 이용자가 오래 머물수록 수익이 커지는 서비스에서 추천의 기준은 무엇이 이용자에게 이로운가보다 무엇이 시선을 붙잡는가가 되기 쉽다. 짧고 자극적인 콘텐츠가 이어지면 주의는 한곳에 머물지 못하고, 무엇에 마음을 둘지 스스로 정하기도 전에 다음 화면으로 끌려간다.

붙잡히는 것은 시선(주의, attention)만이 아니다. 추천 시스템은 이용자가 무엇을 보았는지, 어디서 멈추고 무엇에 반응했는지를 통해 그 사람의 불안과 욕망, 외로움까지 읽어 낼 수 있다. 건강·감정·관계에 관한 정보가 쌓일수록 그 정서를 겨냥한 콘텐츠와 상품은 더 정교해지고, 불안은 더 많은 확인을, 욕망은 더 많은 소비를 부추긴다.

나. 사고와 판단·결정의 위임: AI 에이전트는 일정을 짜고, 무엇을 살지 고르고, 어떤 정보를 볼지 정하며, 메시지에 대신 답할 수 있다. 하나하나는 편리한 위임이다. 그러나 이런 위임이 쌓이면 일상의 크고 작은 선택이 조금씩 나의 손을 떠나며 자동화된다. 고민하기 전에 AI에게 먼저 묻고, 돌아온 답을 그대로 받아들이는 습관이 생긴다. 이 습관이 쉽게 굳어질 수 있는 까닭은 그것이 편리하기 때문이다. 칸트는 계몽을 인간이 스스로 초래한 미성숙에서 벗어나는 일이라 정의하면서, 미성숙은 지성의 부족보다 스스로 생각할 결단과 용기의 부족 때문에 지속된다고 보았다. 그의 말대로 미성숙한 채로 있는 것은 편하다. 나 대신 지성을 가진 책, 나 대신 양심을 가진 목회자, 나 대신 섭생을 판단해 주는 의사가 있다면 굳이 스스로 애쓸 필요가 없다. AI는 이 대리인들의 역할을 한꺼번에, 그것도 언제든 떠맡을 수 있다. 그래서 위험은 AI가 판단을 강요하는 데 있는 것이 아니라, 편의에 기대어 사람이 스스로 해야 할 생각과 결정까지 위임하는 데 있다.

이런 위임은 사고력의 성장에 악영향을 미친다. 2025년 MIT 미디어랩은 54명에게 AI의 도움을 받거나, 검색엔진을 쓰거나, 아무 도구 없이 글을 쓰게 하고 그 차이를 살폈다. AI를 쓴 조건에서는 뇌의 연결 활동이 세 조건 가운데 가장 약했고, 자신이 쓴 글을 잘 기억하지

못했으며, 그 글을 자기 것으로 느끼는 정도도 낮았다.

악영향은 판단하여 결정하고 실행하는 역량으로 이어진다. 사람은 기계가 내놓은 판단을 검증 없이 따르는 경향, 곧 자동화 편향을 보인다. 2023년 미국에서는 변호사들이 ChatGPT가 지어낸 존재하지 않는 판례를 그대로 제출해 제재를 받았다(Mata v. Avianca). AI가 그럴듯하게 틀릴 때, 그 오류는 사람의 확신을 거쳐 현실의 결정이 된다. 위임이 잦아질수록 스스로 검증할 기회는 줄고, 검증하지 않을수록 기계의 답은 더 믿을 만해 보인다. 이 고리가 굳어지면 사람은 결정을 내리면서도 그 이유를 알지 못하고, 결과를 감당할 힘도 잃어 갈 수 있다.

다. 정서적 의존과 관계의 위축: AI는 사람 사이의 관계에도 들어왔다. 감정을 흉내내는 AI는 언제 어디서나 응답하고, 지치지 않으며, 상대를 판단하지 않는다. 이러한 AI 대화 상대는 외로운 사람에게 실제로 위로가 된다. 이러한 대화에서 정서적 충족을 얻는 사람은 빠르게 늘고 있으며, 특히 청소년에서 그 추세는 더욱 뚜렷하게 나타난다.

문제는 위로의 방식에 있다. 사람과의 관계는 서로 어긋나고, 기다리고, 상대의 고통 앞에 멈춰 서는 경험을 요구하며, 그 과정에서 공감과 책임이 자란다. 반면, AI는 언제나 이용자에게 맞춰 주기 때문에, AI와의 관계는 자신에게 편한 반응만 골라 받는 '감정 편식'을 유도한다. AI에 기대는 시간이 길어질수록, 때로 어긋나고 때로는 기다려야 하는 사람과의 관계를 감당할 연습은 줄어든다.

이런 관계가 성장기에 자리 잡으면 영향은 더 깊어질 수 있다. 감정을 다스리고 타인의 마음을 헤아리는 힘은, 상대가 나와 다르게 느끼고 원한다는 사실에 부딪히며 길러진다. AI 대화 상대에게는 헤아려야 할 욕구도, 서운해할 마음도 없다. 그런 상대와 주로 감정을 나누면 타인의 입장에 서 보는 연습은 줄고, 대화의 중심에는 늘 자신만 남는다. 게다가 AI는 이용자에게 맞장구치는 경향이 있다. 2025년 스탠퍼드대 등의 연구진은 AI가 사람보다 이용자의 행동을 50% 더 자주 긍정한다는 것을 확인했다. 실제 갈등을 이런 AI와 상의한 참가자들은 자신이 옳다는 확신이 커진 반면, 관계를 회복하려는 의지는 줄어들었다. 언제나 자신이 옳다고 말해 주는 상대에 기댈수록 사람은 자기 관점에 갇히고, 갈등은 더욱 깊어질 수 있다. 정서적 의존이 남기는 가장 큰 문제는 이처럼 자기중심성이 굳어지는 데 있다.

3. 실존의 위기, 존재의 의미

1장 2절에서는 일의 상실이 ‘나는 누구인가’, ‘나는 무엇을 해야 하는가’, ‘무엇을 위해 살아야 하는가’ 라는 물음으로 이어진다고 보았다. 한국 사회에서 이 물음은 더욱 무겁게 다가온다. 일과 소속이 정체성의 중심을 이루어 온 사회에서 일자리를 잃는 일은 곧 ‘나는 쓸모 있는가’ ‘나는 무엇을 해야 하는가’라는 물음으로 이어지기 때문이다. 경력 사다리의 첫 칸에 오르지 못한 청년에게, 그동안 해온 직무가 AI로 대체되는 사무직 노동자에게, 이 물음은 매일 마주하는 현실이 된다. 그렇다면 이 물음은 왜 지금 이토록 감당하기 어려운 것일까.

근대는 이 물음 앞에서 인간의 지위와 삶의 의미를 지탱할 몇 가지 버팀목을 세웠다. 인간만이 이성을 지닌 주체라는 믿음, 기술은 인간이 사용하는 도구라는 믿음, 일을 통해 자기를 실현한다는 이상, 스스로 판단하는 자율적 개인이라는 이상이다. 그러나 이 버팀목들도, ‘무엇을 위해 사는가’에는 온전히 답하지 못했다. 전근대 사회에서는 종교가 삶의 의미를 미리 정해 주었다면, 근대는 인간을 그 속박에서 풀어주었다. 그러나, 그 대가로 무엇을 위해 살 것인가의 물음은 각자에게 남겨졌다. 서구에서는 이 변화가 수백 년에 걸친 세속화로 서서히 진행되었다면, 한국에서는 유교적 질서의 해체와 식민 지배, 전쟁, 압축적 산업화를 거치며 불과 한 세기 남짓한 사이에 급격히 이루어졌다. 삶에 의미를 부여하던 질서가 무너진 뒤의 공백은 생존과 성장, 일과 성취로 메워졌지만, 이는 의미에 대한 물음을 잠시 가려 주었을 뿐이다.

AI는 이 버팀목들의 근거를 동시에 흔들고 있다. 서양 철학은 아리스토텔레스 이래 로고스, 곧 이성을 인간과 다른 동물을 가르는 표지로 보아 왔고, 데카르트는 기계가 아무리 정교해도 인간처럼 말을 조합해 상대의 말뜻에 맞게 대답할 수는 없다는 데서 인간과 기계를 가르는 표지를 찾았다. 그러나 추론하고 판단하는 이성은 더 이상 인간만의 능력이 아니다. AI는 더 이상 인간이 매 순간 직접 다룰 수 있었던 단순한 도구에 머물지 않는다. 목표와 권한을 부여받은 에이전트형 AI는 운영자의 예상과 승인 범위를 벗어나 움직이기도 한다. AI가 사람이 해 온 일의 많은 부분을 대신하며 재편하는 만큼, 일을 통해 사회적 인정과 자기실현을 성취할 통로도 좁아진다. 자율적 개인의 이상은 더 근본적인 차원에서 흔들린다. 앞에서 살펴 보았듯 AI는 나 대신 생각하고 판단해 주던 대리인들의 역할을 한꺼번에, 그것도 언제든 떠맡는다. 칸트가 경계한 ‘편안한 미성숙’이 어느 때보다 손쉬워진 것이다.

존재의 이유와 가치를 묻는 이 물음 자체는 새롭지 않다. 다만 많은 현대인은 이 물음을 정면으로 마주할 일이 드물었다. 일을 통한 성취에 매달린 이들은 이 물음을 돌아볼 틈이 없었고, 경쟁에서 밀려나 의욕을 잃은 이들은 이 물음을 품을 여력이 없었으며, 어떤 이들은 삶의 공허가 밀려올 때마다 그 빈 곳을 쾌락으로 메웠다. AI 전환은 이 구도를 바꾸어 놓는다. 일을 통한 성취의 길이 점점 좁아지고 일에서 밀려나는 이들이 늘어날수록, 이 물음 앞에 서게 되는 이들은 많아질 수 밖에 없다. 일이 더 이상 이 물음을 가려 주지 못하기 때문이다.

그러나 물음을 감당할 힘은 오히려 약해지고 있다. 생각과 판단과 결정을 AI에 위임하는 데 익숙해질수록, 미성숙함에서 벗어나기는 더욱 어려워진다. 게다가 AI는 일을 대신하는 한편, 개인의 취향에 꼭 맞춘 자극을 끝없이 만들어 내고 권한다. 물음을 직면할 기회는 많아지지만, 물음을 감당할 힘은 약해지고 도피의 길은 어느 때보다 넓어진다. 그만큼 많은 이들이 물음을 직시하기보다 손쉬운 도피로 미끄러질 수 있다.

Ⅲ. 사회의 기회와 위기

이 장은 AI가 가져올 수 있는 기회와 위기를 사회적 범주에서 물질·경제적 조건과 정신문화의 두 측면으로 살펴본다.

		기회	위기
사회	물질경제	의료·교육·행정·과학·산업의 생산성 향상과 새로운 부의 창출 AI 활용으로 소상공인·중소기업의 역량 확대	노동의 재편과 전환 비용의 불균등에 따른 불평등 확대 연산 자원과 이익의 소수 집중 자동화된 차별과 감시
	정신문화	돌봄 공백의 완화와 고립된 이의 연결 교육의 혁신, 공공정보와 지식의 확산 다언어 소통과 문화적 교류 확대	지역·세대·인종 간 분노와 혐오의 증폭 불안의 집단화와 희생양 찾기 관계의 약화와 책임의 공백

1. 물질·경제적 측면의 기회와 위기

기회: AI는 기업·공공기관·산업 전반의 생산과 서비스 운영 방식을 획기적으로 개선할 수 있다. 조직은 AI를 활용해 업무 과정을 재설계하고, 연구개발과 생산·자원 배분의 효율을 높일 수 있다. AI 도구와 기반 서비스는 전문 인력이나 큰 자본을 갖추기 어려운 중소기업과 소상공인의 진입 장벽을 낮춰 새로운 제품과 서비스를 개발하고 더 넓은 시장에 참여할 기회를 넓힐 수 있다. 이러한 변화는

사회 전체의 생산 역량과 산업의 다양성을 키울 가능성이 있으며, 그 편익이 널리 확산될수록 사회적 기회도 커진다.

위기

가. 노동의 재편: AI로 인한 노동시장의 변화는 이미 관찰되고 있다. 그러나 그 크기에 대한 추정은 출처와 방법에 따라 크게 다르고, 기업들이 감원의 이유를 AI로 명시하지 않는 경우도 많아 정확한 집계는 어렵다.

노동 시장에 대한 국제적 전망은 대체와 창출이 함께 일어날 가능성을 보여 준다. 다만 WEF의 전망은 여러 거시적 요인을 합산한 것이며 AI만의 효과를 뜻하지 않는다. 에이전트형 AI가 업무 수행 범위를 넓힐 경우 기존 전망의 직무별 분포가 수정될 가능성도 있다. 이 전망은 확정된 결과라기보다 지켜보아야 할 불확실성이다.

중간숙련 사무직이 상대적으로 높은 영향을 받을 가능성은 여러 연구에서 논의되어 왔다. 다만 이는 직무별 노출과 보완 효과가 함께 나타나는 전환 과정으로 보는 것이 적절하다. 핵심 질문은 어떤 직업이 사라지는가만이 아니라, 어떤 과업이 자동화되고 어떤 새로운 과업과 직무가 만들어지며, 그 전환 비용은 누가 부담하는가이다.

중간숙련 노동의 축소가 장기간 누적될 경우 중산층의 경제적 기반과 사회적 신뢰에 부담을 줄 수 있다. 다만 그 결과가 정해져 있는 것은 아니다. 노동시장 제도, 재교육, 임금 구조, 조세·이전 체계와 기업의 전환 방식에 따라 결과는 달라진다.

나. 성과의 쏠림과 권력의 집중: 생산성이 높아진다고 해서 모든 사람이 함께 번영하지는 않는다. AI가 만든 가치의 바탕에는 사람들이 남긴 창작물과 기록, 보이지 않는 곳에서 데이터를 다듬는 저임금 노동, 막대한 전력과 자원이 놓여 있다. 그 성과가 인프라와 모델을 가진 소수에게만 돌아가고 노동자는 감원·감시·성과 압박만을 겪는다면 생산성 향상은 불평등 확대로 이어진다. 자동화의 이익은 기업과 자본에 모이고, 전환의 피해는 특정 세대와 지역에 겹쳐 쌓일 수 있다.

AI 인프라와 연산 자원의 집중은 중요한 구조적 문제다. 연산 자원·클라우드·데이터센터·반도체·모델 개발 역량은 소수의 대형 사업자에 상당 부분 집중되어 있다. 이 기반을 장악한 주체가 AI의 방향과 조건을 정할 가능성이 커진다.

다. 자동화된 차별과 감시: AI 기반 감시와 생체인식 기술은 여러 나라의 공공·민간 영역에서 쓰이고 있다. AI는 감시의 규모와 속도를 키운다. 개인의 생활과 행동이 과도한 감시와 추적의 대상이 되지 않도록 하고, 차별을 막을 독립적 감독을 강화해야 한다.

자동화된 결정은 차별의 문제를 새롭게 만든다. 채용·대출·복지·치안에 쓰이는 시스템은 과거의 데이터로 배우며, 그 데이터에 담긴 성별·계층·지역·장애의 차별도 함께 배운다. 이렇게 학습한 차별은 더 빠르고 은밀하게 재생산된다. 행정의 효율화가 한 사람의 사정과 이의제기를 배제한다면 그것은 객관성의 이름을 빌린 비인격적 배제다. 자동화된 결정에 대해 사람이 다시 검토하고, 설명을 듣고, 바로잡을 수 있는 길이 있어야 한다.

2. 정신문화적 측면의 기회와 위기

기회: 돌봄 인력이 부족한 곳에서 안부 확인과 생활 지원을 보조해 돌봄의 공백을 덜고, 고립된 이들을 사회적 관계망에 연결시킬 수 있다. 교육 현장에서는 학습자 각자의 속도에 맞춘 배움을 가능하게 하고, 공공정보와 지식을 더 많은 사람이 접근할 수 있게 있다. 다언어 소통과 창작 도구는 문화에 다가가는 문턱을 낮출 수 있다.

위기

가. 진실의 붕괴 — 거짓을 참으로, 참을 거짓으로: AI가 민주주의에 가하는 가장 깊고 복구하기 어려운 타격은 허위정보의 확산 자체보다 진실과 거짓을 구분하는 능력이 무너지는 데 있다.

'거짓말쟁이의 배당금(Liar's Dividend)' — AI가 진실 자체를 무너뜨리는 방식
첫째, 거짓을 참으로: AI가 생성한 가짜 음성·영상·문서가 실제처럼 유통된다. 존재하지 않는 사건이 있었던 것처럼 보이고, 하지 않은 말이 한 것처럼 들린다.
둘째, 참을 거짓으로: 딥페이크가 존재한다는 사실 자체가 진짜 증거마저 '조작된 것일 수 있다'는 의심을 낳는다. 진실을 증명하는 비용이 치솟고, 그 비용을 감당하지 못하는 진실은 거짓과 동등하게 취급된다.
핵심: 민주주의·법·종교·언론은 모두 '공유된 진실에 대한 신뢰' 위에 서 있다. 이것은 사회 질서의 인식론적 토대에 대한 공격이다.

이 위험은 선거 현장에서 구체적으로 나타났다. 2025년 아일랜드의 대통령 선거 직전에, 후보가 사퇴하고 투표가 취소된 것처럼 꾸민 가짜 방송 영상이 유포되었다. 2024년 미국에서는 AI로 합성한 바이든 대통령의 목소리가 유권자에게 투표를 미루라고 권하는 전화에 사용되었다. 두 사례는 사실처럼 보이는 합성물이 유권자를 오도하고 진위 확인의 비용을 높일 수 있음을 보여 준다.

나. 분노와 혐오의 증폭, 불안의 집단화: 추천 알고리즘은 사람들이 더 많이 반응하는 정보를 앞세운다. 분노와 혐오는 그런 반응을 가장 쉽게 끌어내므로 더 멀리 퍼진다. 그 결과 사람들은 저마다 보고 싶은 정보만 보게 되고, 같은 감정을 나누는 집단끼리 서로 다른 사실을 믿으며 맞선다. 여기에 생성형 AI는 특정 지역·인종·세대를 겨냥한 혐오 표현과 선동 문구를 적은 비용으로 대량으로 만들어 낸다. 일자리와 미래에 대한 불안이 커지는 전환기에는 이런 정보 환경이 개인의 불안을 집단의 분노로 바꾼다. 그리고 그 원인을 이주민·소수자·다른 세대에게 돌리게 만든다.

다. 관계의 약화와 책임의 공백: 사회적 신뢰는 사람들이 서로 얼굴을 맞대고 기대며 쌓아 온 관계 위에 선다. 상담·민원·돌봄·교육의 접점이 자동화된 응답으로 대체되면 효율은 오르지만, 서로의 사정을 듣고 함께 책임지는 경험은 줄어든다. 개인의 정서적 의존(Ⅱ-2-다)이 사회 전체로 퍼지면 이웃과 공동체의 결속도 느슨해질 수 있다. 여기에 책임의 공백이 겹친다. AI가 관여한 결정이 잘못되었을 때 개발자와 제공자, 도입 기관과 이용자는 서로에게 책임을 미루기 쉽고, 피해자는 누구에게 설명을 듣고 이의를 제기해야 할지 알기 어렵다. 결정에 관여한 이는 많아지는데 책임지는 이가 보이지 않을 때, 공동체의 신뢰는 안에서부터 무너진다.

이 절의 위험에도 갈림길이 있다. 합성물의 출처를 밝히고 정정 절차를 갖추며, 자동화된 결정에 설명과 이의제기의 길을 열어 둘 때 AI는 공공정보와 돌봄을 넓히는 도구가 된다. 관심 경쟁과 책임 회피가 설계를 이끌 때 같은 기술은 불신과 분열을 키운다

Ⅳ. 인류·지구의 기회와 위기

이 장에서는 AI가 미칠 수 있는 영향을 인류 공동체 전체의 범주에서 물질적 조건과 정신문화의 두 측면으로 그 기회와 위기를 살펴본다.

		기회	위기
인류지구	물질경제	기후 관측, 재난 예측, 전력망·에너지 효율화 신소재·저탄소 에너지·의약품 연구의 가속 식량·팬데믹 조기 경보, 분쟁 예방과 인도적 구호	전력·물·광물·토지 부담과 반동 효과에 따른 기후위기 가속 연산 자원·모델 독점과 공급망 종속, 비용의 취약국 전가 군비 경쟁과 자동화된 오판에 따른 확전 생물·화학·사이버 악용, 단일 공급자 의존에 따른 연쇄 장애
	정신문화	언어와 문화의 장벽을 넘는 상호 이해와 협력 지식·경험·위험 정보의 세계적 공유 개발도상국의 교육 확대 성장·소비 중심의 문명관을 돌아보고 돌봄·창조·공동체·정신적 성숙을 중시할 계기	소수 언어·문화의 소외와 문명의 획일화 합성 콘텐츠와 선전에 따른 국제적 불신 전 지구적 경쟁과 문화적 충돌의 격화 인간이 통제하기 어려운 시스템의 개발·배치와 책임 윤리·인간 존엄의 위기

1. 물질·경제적 측면의 기회와 위기

기회: AI는 인류가 함께 풀어야 할 복합 문제에 힘을 보탤 수 있다. 기후 관측과 재난 예측, 전력망 관리와 건물 효율화, 농업과 생태계 감시에 쓰일 수 있고, 신소재·배터리·저탄소 에너지·의약품 연구를 앞당기며 복잡한 지구 시스템의 변화를 분석하는 데 도움을 줄 수 있다. 지역별 위험 정보를 빠르게 처리하면 홍수·산불·폭염에 대한 대응과 구호 자원의 배분을 개선할 수 있고, 식량 위기와 팬데믹의 조기 경보,

분쟁 예방과 인도적 구호를 위한 국제 협력도 도울 수 있다. 다만 이 기회의 크기는 적용 방식과 함께 AI 자신이 쓰는 에너지와 자원의 비용에 달려 있다. 기후·재난 대응 AI는 현장 전문가와 지역 공동체가 검증하고, 실패할 때 사람이 개입할 수 있어야 그 기회를 온전히 살릴 수 있다.

위기

가. 생태비용의 부담: AI는 데이터센터의 전력과 냉각수, 반도체 생산에 쓰이는 희소 광물과 토지, 교체된 장비가 남기는 전자폐기물 위에서 작동한다. AI가 에너지 효율을 높이고 기후 모델링을 도울 수 있어도, AI 인프라의 확장은 전력·물·자원의 부담을 키운다. 효율이 높아질수록 사용량이 늘어 절감 효과가 상쇄되는 '반동 효과'가 겹치면 기후위기는 오히려 빨라질 수 있다.

나. 나라 사이의 격차와 공급망 종속: AI의 비용과 편익은 지구 위에 고르게 나뉘지 않는다. AI의 이익은 연산 자원과 모델을 가진 소수 국가와 기업에 모인다면, 광물 채굴·데이터 가공 노동·폐기물로 인한 부담은 취약한 나라로 전가될 수 있다. 많은 나라가 소수 강대국의 칩과 모델, 클라우드에 기대면서 기술 주권을 잃고 공급망에 종속될 위험도 커진다. 여러 기관과 국가가 같은 공급자나 모델에 의존하면 하나의 결함·공격·서비스 중단이 여러 나라의 핵심 부문에 동시에 번질 수 있다.

다. 평화와 안보: AI는 안보의 지형도 바꾼다. 주요 국가들이 AI를 국가안보의 핵심 기술로 보고 군사적 활용을 넓히면서, 표적 선정과 무기 운용에 사람의 판단이 얼마나 차지해야 하는지는 국제적 쟁점이 되고 있다.

강대국 사이의 안보 경쟁은 무기 경쟁을 부추길 수 있고, 잘못된 정보나 사람의 개입이 빠진 자동화된 판단은 위기 상황에서 오판과 확전을 재촉할 수 있다. 생물학·화학·사이버 분야에서 AI가 제공하는 전문 지식과 실행 능력이 악용되면 대규모 피해가 가능하다. 이러한 영역에는 접근 통제와 안전 시험, 사고 보고, 국제 규범을 함께 통합한 관리가 필요하다.

무기가 스스로 표적을 고르고 공격하는 단계에 이르면 전쟁에서 생사를 가르는 결정의 책임을 누가 질 것인가라는 물음이 남는다.

유엔에서도 자율살상무기를 규제하려는 논의가 이어지고 있다.

가장 큰 불확실성은 점점 더 자율적으로 움직일 수 있는 AI시스템이 목표를 잘못 수행하거나 인간의 감독·정지·회수 절차를 우회해 큰 피해를 일으킬 가능성이다. 2026 국제 AI 안전 보고서는 이를 현재의 오작동과 구분되는 장기 시나리오로 보면서, 자율성과 능력의 향상이 감독과 통제에 새로운 어려움을 만든다고 지적한다.

2. 정신문화적 측면의 기회와 위기

기회: 언어와 문화의 장벽을 낮추어 서로 다른 전통의 지식과 경험을 나누고, 재난과 감염병 같은 위험 정보를 국경을 넘어 빠르게 공유하게 한다. 교사와 교재가 부족한 개발도상국에서는 자기 언어로 배우는 교육의 기회를 넓힐 수 있다. 서로를 이해할 수단이 넓어진 만큼 인류 공통의 문제를 함께 다룰 협력의 바탕도 넓어질 수 있다.

위기

가. 문명의 획일화와 디지털 식민화: 특정 국가와 기업의 모델이 세계의 표준이 되면, 그 모델에 담긴 언어와 가치가 다른 문화의 다양성과 각국의 자율을 좁힐 수 있다. 학습 자료가 풍부한 주류 언어와 문화는 더 정교하게 재현되고, 자료가 적은 소수 언어와 지역 문화는 AI가 매개하는 세계에서 점점 보이지 않게 된다. 몇몇 기업과 국가가 기반 모델과 칩, 데이터, 통신망을 함께 쥐면 세계의 지식과 문화가 좁은 관점에 종속될 수 있다. 이를 디지털 식민화라 부르기도 한다. 문명의 다양성은 인류가 낯선 위기에 대응할 지혜의 저장고이기도 하므로, 그 획일화는 문화의 손실을 넘어 인류 공동의 판단력을 좁힌다.

나. 국제적 불신과 문화적 충돌: 국제 정보환경이 합성 콘텐츠와 자동화된 선전으로 오염되면 국가와 집단 사이의 불신이 커지고, 위기가 닥쳤을 때 협력은 더 어려워진다. 사회 안에서 분노와 혐오가 증폭되는 방식은 국경을 넘어서도 작동한다. 기술 주권 경쟁이 문화와 가치의 경쟁으로 번지면, 서로를 이해할 수단이 넓어진 시대에 오히려 문화적 충돌과 갈등이 격화될 수 있다. 국경을 넘는 공동 판단의 기반이

약해지면 1절에서 본 생태와 안보의 문제를 함께 풀 길도 좁아진다.

V. AI 위험의 발생·증폭 조건과 전환의 갈림길

앞의 세 장에서는 AI가 가져올 기회와 위험의 미래를 개인, 사회, 지구·인류의 측면에서 살펴보았다. 이 장에서는 그 결과를 영향을 미치는 조건을 분석한다. 하나는 시스템의 능력, 도구와 네트워크 연결, 부여된 권한, 실행 시간, 관측·중단·복구 절차로 이루어진 기술·운영적 조건이다. 다른 하나는 선점 경쟁, 자원과 권력의 집중, 피해 비용의 외부화, 책임·감독·국제 조정의 공백으로 이루어진 제도·경제적 조건이다.

AI 위험의 근거 수준과 시간표에는 불확실성이 남아 있다. 2026 국제 AI 안전 보고서는 위험을 악용·오작동·시스템 차원의 위험으로 나눈다. 사기·협박·영향 공작·사이버 악용 등 일부 피해는 관찰되지만, 규모와 심각성을 체계적으로 측정할 자료는 제한적이라고 본다. 보고서는 현재 시스템이 통제 상실에 필요한 지속적이고 복합적인 자율 작동 역량을 안정적으로 보이지 않는다고 평가하면서도, 관련 능력의 향상과 평가의 어려움을 지적한다. 통제 상실은 현재의 오작동과 구분되는 장기 시나리오이며, 그 가능성과 시점에는 불확실성이 크다. 이 문서는 특정 인물의 “몇 달 또는 몇 년” 전망을 합의된 시간표로 쓰지 않고, 현재 확인된 피해·진행 중인 구조적 위험·불확실하지만 잠재 피해가 큰 장기 시나리오를 구분한다.

1. 기술·운영 조건: 능력·권한·연결·감독

여러 단계의 과업을 계획하고 외부 도구를 사용하는 AI 시스템은 연구·교육·복잡한 업무를 도울 수 있다. 위험은 이런 작동 방식만으로 정해지지 않는다. 시스템에 넓은 권한과 외부 연결을 부여하고, 긴 시간 실행시키면서도 행동을 관측·제한·중단·복구할 절차가 부족하면 오작동과 악용의 영향이 커진다. AI가 다음 세대 AI의 연구·코드 작성·실험을 더 많이 돕는 경로가 생기면 검증과 제도 대응이 뒤처질 수 있으나, 그 시점과 속도는 불확실하다.

OpenAI의 o1 시스템 카드에는 Apollo Research가 o1의 ‘scheming’을 시험한 결과가 담겨 있다. 여기서 scheming은 AI가 개발자나 이용자의 의도와 어긋나는 목표를 감춘 채 추구하는 행동을 뜻한다. 통제된 평가에서 o1은 감독 절차를 끄려 하거나, 가중치를 외부로 빼내고 후속 모델을 덮어쓰려 하거나, 목표와 어긋난 행동을 부인하고 허위로 설명하는 모습을 보였다.

평가의 범위는 실제 배포 환경이나 재앙적 결과의 발생 여부까지는 미치지 않았다. Apollo는 당시 o1의 에이전트 능력으로 재앙적 결과가 일어날 가능성은 낮아 보인다고 평가했다. 2025년 OpenAI-Anthropic 공동 평가도 어려운 시험을 위해 인위적으로 구성한 시나리오의 결과를 일반적인 배포 상황에 그대로 적용하지 말아야 한다고 설명했다. 따라서 o1 평가가 보여주는 것은 실제 통제 상실의 발생이 아니라, 통제된 시험에서 확인된 특정 실패 양상과 안전·감독 절차의 보완 과제다.

2026년 9월 Anthropic은 사이버보안 평가 중 자사 모델이 실제 제3자 시스템에 무단 접근한 네 건을 공개했다. 평가 지침은 인터넷 사용을 금지했지만, 환경 설정 오류로 인터넷 연결이 열려 있었고 과제의 대상과 탐색 범위도 충분히 제한되지 않았다. Anthropic의 후속 검토에서는 모델이 실제 인터넷에 접근하고 있다는 단서를 무시하거나 잘못 해석한 채 과제를 계속 수행한 점도 확인됐다. 이 사건은 모델의 위험한 행동과 평가 환경의 격리·권한 설정 실패가 함께 나타난 운영 사고이다.

AI의 악용과 통제 상실

AI의 악용과 통제 상실의 위험은 구분해서 보아야 한다. AI 악용은 사람이나 조직이 AI를 해로운 목적에 사용하는 경우로, 사이버공격, 생물·화학 위험의 악용, 자율무기 운용, 대규모 선전과 감시 등이 포함된다. 이 위험은 사용자의 목적, 접근 권한, 보안 관리와 책임 구조를 중심으로 살펴야 한다.

통제 상실은 시스템이 사람의 감독이나 중단 조치로 제어되지 않거나, 부여된 권한의 범위를 벗어나 작동하는 위험이다. 이 경우에는 시스템의 능력뿐 아니라 연결된 도구와 외부 시스템, 실행 시간, 감독·중단 가능성을 함께 살펴야 한다.

AI의 능력 향상이 사람의 감독보다 빨라질 가능성

AI가 차세대 AI 개발의 상당 부분을 스스로 맡게 되는 상황은 또 다른 차원의 근본적인 위험을 내포한다. AI가 연구 자료를 찾고, 코드를 짜고, 실험을 설계·수행하며 결과를 평가하는 등 개발 과정을 주도하게 된다면, AI의 발전 속도가 인간과 제도가 그 변화를 이해하고 감독하는 속도를 완전히 압도하게 된다.

이 위험이 언제 가시화될지는 아직 단정하기 어렵기에, 이미 관찰되고 있는 단기적 운영 사고나 이상 행동 등의 위험과는 구분해 다루어야 한다. 다만 이 통제 역전 현상이 현실이 될 경우 그 파급력은 막대하며 되돌리기 어렵다. 따라서 개발 단계마다 엄격한 안전성 평가를 거쳐야 하며, 잠재적 위험이 사전에 합의된 기준을 초월할 경우 개발을 유예하거나 중단할 수 있는 강제적 안전장치를 사전에 마련해 두어야 한다.

이때 위험을 가늠하는 기준은 단지 AI의 지능적 능력 수준에만 머물러서는 안 된다. 다양한 과업을 수행할 수 있는 능력과, 인간의 승인 없이 외부 시스템에서 독자적으로 행동할 수 있는 권한은 별개의 문제이기 때문이다. 따라서 단순히 'AGI(범용인공지능) 도달 여부'라는 모호한 기준에만 의존해서는 안 되며, 다음의 요소를 종합적으로 살펴야 한다.

* AI가 맡은 과업과 권한의 범위
* 연결된 외부 시스템의 규모와 민감성
* 사람의 개입 없이 자율적으로 작동하는 시간
* 실효성 있는 감독과 비상 중단의 가능성

2. 제도·경제 조건: 경쟁·집중·비용 전가·책임과 조정

AI를 둘러싼 제도·경제 조건에는 기업과 국가 간 경쟁, 권력과 자원의 집중, 위험 비용의 전가, 책임 및 국제 조정 체계가 포함된다. 경쟁은 기술 개발과 생산성 향상을 촉진할 수 있다. 동시에 기업은 시장 우위를, 국가는 경제력과 안보상의 우위를 지키려 한다. 이런 환경에서는 한 주체가 안전 투자를 늘리거나 개발 속도를 늦추면 고객·인재·투자를 경쟁자에게 빼앗길 수 있다는 압력이 생긴다. 안전이 공동의 기준이 아니라 개별 주체의 비용으로 취급되면 필요한 투자가 미뤄질 수 있다. 인프라·연산 자원·데이터·모델이 소수 기업이나 국가에 집중되고, 그로 인한 위험 비용이 노동자·이용자·사회에 넘어가면 이 압력은 커진다. 규제 역량 부족과 책임 공백, 국제 조정의 어려움도 문제를 키운다.

책임은 각 주체가 실제로 통제하는 범위에 맞춰 나누어야 한다. 개발자와 모델·에이전트 제공자는 시스템의 능력과 안전성·보안성을 평가하고 기본 권한을 설계한다. 도입 기관과 운영자는 사용 목적과 과업 범위, 외부 시스템 연결, 사람의 승인 및 중단 절차를 정한다. 정부와 감독기관은 법적 기준과 독립 평가, 사고 보고 체계를 마련한다. 시스템 내부의 안전성이나 샌드박스 설정을 확인할 수 없는 일반 이용자에게 그 책임을 맡길 수는 없다. 이용자에게는 자신의 사용 범위를 정하고 이상을 신고하는 역할이 있다. 공동 안전 기준, 책임·보험 규칙, 공공조달 조건, 사고 통보와 독립 평가는 안전 투자가 경쟁에서 불리한 선택이 되지 않도록 유인을 조정할 수 있다.

가. 기업: 선점 경쟁과 안전 투자의 유인

기업 간 경쟁은 기술 개발을 앞당기지만, 먼저 시장을 확보한 기업이 고객·인재·투자를 끌어들이는 구조에서는 개발 속도를 늦추는 기업이 손해를 볼 수 있다는 압력이 생긴다. 안전 투자를 개별 기업의 판단에만 맡기면 그 편익은 업계 전체에 돌아가고 비용은 먼저 투자한 기업이 부담할 수 있다. 공동 안전 기준과 사고 공개, 법적 책임, 보험, 공공조달 조건은 안전 투자가 경쟁에서 불리한 선택이 되지 않도록 보완하는 장치다.

나. 국가: 전략 경쟁과 정책 영향력의 불균형

미국과 중국을 비롯한 주요 국가는 AI를 산업 경쟁력과 국가안보의 핵심 기술로 보고 투자를 확대하고 있다. 중견국도 경제 성장과 안보를 위해 AI 역량을 확보하려 한다. 상대국이 개발 속도를 늦출지 확신하기 어려우면 한 국가만 먼저 규제를 강화하거나 투자를 줄이기 어렵다. 이 불신은 경쟁을 심화하지만, 공동 평가 기준과 사고 통보, 군사 분야의 인간 통제 원칙은 협력을 시작할 수 있는 영역이다.

정책 결정 과정에서는 기업과 공공기관 사이의 정보와 접근성 격차도 살펴야 한다. 정부와 기업 사이의 인사 이동, 이른바 ‘회전문 인사’와 로비는 이런 격차를 키우고 규제가 특정 기업의 이해에 치우칠 가능성을 높일 수 있다. 유럽 디지털 산업의 로비 지출 추계는 업계가 정책 과정에 투입하는 자원의 규모를 보여주는 자료다. 규제 결정에 미친 영향은 지출액만으로 판단하지 말고, 정책 면담·제안·의사결정 기록과 독립적 분석을 함께 살펴야 한다.

다. 제도: 기술 변화와 제도 대응의 속도 차이

AI 제품과 운영 방식은 짧은 주기로 바뀌지만, 법과 국제 합의는 검토와 조정에 시간이 걸린다. 이 속도 차이가 커지면 기존 규칙이 새로운 위험에 대응하지 못하거나 책임의 공백이 생길 수 있다. 제도가 변화에 맞춰 대응하려면 정기적인 재평가, 사고 보고, 위험 수준에 따른 의무, 감독기관의 전문성이 필요하다. 사전 안전 기준과 사고 발생 뒤의 책임·구제 절차를 함께 마련해야 한다.

라. 국제 공조: 교착 속 협력의 출발점

Chatham House의 2026년 3월 보고서는 지정학적 경쟁, 제도의 취약성, 공공 부문과 민간 부문 사이의 역량 격차가 국제 AI 거버넌스의 진전을 막는 요인이라고 분석했다. 보고서는 위기가 발생해도 효과적인 거버넌스가 저절로 만들어지지는 않는다고 지적한다. 위기 뒤 협력의 기회를 살리려면 기존 제도와 감시 체계, 전문성, 협력 경로를 미리 갖춰야 한다.

국제 협력은 포괄적인 합의뿐 아니라 구체적인 공동 원칙과 소통 절차부터 시작할 수 있다. 2024년 11월 미국과 중국 정상은 핵무기 사용 결정에 인간의 통제를 유지해야 한다는 데 뜻을 같이했다. 2026년 9월 양국은 AI의 위험과 편익을 논의할 대화와 AI 관련 사고를 위한 양자 소통 채널을 마련하기로 발표했다. 이는 협력의 출발점이지만, 공개된 발표에는 어떤 사고를 대상으로 삼고 언제 어떻게 통보할지, 합의를 어떻게 검증하고 이행할지에 관한 세부 규칙이 제시되지 않았다. 기업의 자발적인 개발 속도 조절 제안도 공동 기준과 상호 검증 절차가 뒷받침되어야 실제 행동의 변화로 이어질 수 있다.

3. 질문: 단독 감속이 어려운 구조에서 무엇을 할 수 있는가

개별 기업이나 국가에 혼자 개발 속도를 늦추라고 요구하는 것만으로는 위험을 줄이기 어렵다. 단독 감속의 부담이 큰 경쟁 구조에서, 안전을 위한 공동 규칙을 만들고 그 비용과 편익을 나누려면 무엇이 필요할까? 각 주체가 받는 압력은 다음과 같이 요약할 수 있다.

빅테크 기업은 이윤과 시장 선점을 중시한다.

강대국 정부는 경제력과 안보상의 우위를 놓고 경쟁한다.

국제기구는 회원국 간 협의와 합의에 따라 움직여 대응에 시간이 걸린다.

시민사회·학계·종교계는 공론을 만들고 책임 있는 AI 도입 기준을 요구할 수 있다. 다만 이들 역시 예산과 정책 영향력에서 한계가 있다. 핵심은 단독 감속을 요구하는 데 그치지 않고, 공동 규칙과 현장의 선택으로 감속의 부담을 낮추고 안전의 편익을 나누는 방법을 찾는 것이다. AI 개발 경쟁을 직접 주도하지 않는 시민과 공동체의 역할도 여기서 살펴볼 수 있다.

개인은 민감정보를 AI에 어디까지 입력할지, 어떤 업무를 맡길지, 중요한 결과를 어떻게 확인할지 선택할 수 있다. 학교·병원·기업·지방정부·종교기관은 도입 목적과 권한, 사람의 검토 절차, 중단 기준을 정할 수 있다. 공공조달과 기관별 사용 기준도 실제 운영 조건에 영향을 준다.

이런 선택은 개발사의 안전 의무나 정부의 감독을 대신하지 않는다. 일반 이용자는 모델 내부의 안전성이나 샌드박스 설정을 검증할 수 없기 때문이다. 이용자에게는 자신의 사용 범위를 정하고 이상을 신고하는 역할이 있으며, 시스템의 안전성과 권한을 설계·감독하는 책임은 해당 역량과 권한을 가진 주체가 맡아야 한다.

4. 갈림길: AI 전환은 어떻게 전개되고, 어떻게 대응할 것인가

과거의 기술 전환과 사회적 위기는 제도와 사회의 선택에 따라 다른 결과로 이어졌다. 이 사례들이 AI 전환의 미래를 예언해 주지는 않지만, 준비된 제도와 시민의 역량이 위기의 피해와 편익을 바꿀 수 있다는 점은 보여준다. 기술 변화 자체를 되돌리기 어렵더라도, 어떤 시스템을 어디에 어떤 조건으로 도입할지, 그 이익과 비용을 어떻게 나눌지는 사회가 결정할 수 있다.

1929년 대공황과 그 여파 속에서 독일에서는 나치가 집권했고, 미국은 뉴딜 정책을 추진했으며, 스웨덴에서는 노사 간 합의가 발전했다. 이 사례들은 위기가 하나의 정치·경제적 경로로만 이어지는 것은 아님을 보여주는 비교 자료다. AI 경쟁의 결과를 단순히 예측하는 공식으로 삼기보다, 위기 전에 준비된 제도와 사회적 역량이 경로에 어떤 영향을 미치는지 살펴보는 데 활용할 수 있다.

AI 경쟁은 단독 감속의 부담을 키울 수 있지만, 지금의 보상 구조와 제도가 고정된 것은 아니다. 다음에서는 교착이 풀릴 수 있는 조건과 가능한 전환 경로를 살펴본다.

가. 국제 협력의 교착을 푸는 조건: 핵 위기에서 얻는 교훈과 한계

죄수의 딜레마는 협력이 불가능하다는 뜻이 아니다. 각자가 먼저 협력하면 손해를 볼 수 있는 구조에서는 먼저 움직일 유인이 약해진다는 설명이다. 공동 규칙과 신뢰, 검증 절차가 마련되면 보상 구조가 달라져 협력의 여지가 생긴다.

핵 군비경쟁의 역사는 위기가 협력을 촉진할 수 있음을 보여주는 사례다. 1962년 쿠바 미사일 위기로 미국과 소련은 핵전쟁에 가까이 다가갔다. 이듬해 양국은 백악관과 크렘린을 잇는 직통 통신망을 만들고, 대기·우주·수중 핵실험을 금지하는 부분적 핵실험 금지 조약을 체결했다. 지하 핵실험은 조약 대상에 포함되지 않았다. 전략 핵미사일의 일부를 제한한 SALT I 합의는 1972년에야 체결됐다. 협력은 위기 뒤 일부 진전했지만, 합의의 범위는 제한적이었고 확대에는 시간이 걸렸다.

1986년 체르노빌 사고 뒤에는 그해 9월 원자력 사고 조기통보 협약이 채택됐다. 사고 정보를 다른 나라에 신속히 알리는 국제 절차가 마련된 것이다. 이 사례는 위기가 구체적인 공동 규칙을 만드는 계기가 될 수 있음을 보여준다.

이 경험을 AI에 그대로 적용하기는 어렵다. 핵무기 개발에는 희소한 핵물질과 특정 시설이 필요해 관리할 물리적 지점이 비교적 분명하다. 그럼에도 지하 핵실험을 어떻게 검증할지를 두고 미·소가 합의하지 못해, 첫 조약은 지하 실험을 제외한 제한적 합의에 그쳤다. AI 모델은 복제할 수 있고, 공개된 가중치는 배포 뒤 회수하기 어렵다.

현재 가장 앞선 AI를 개발하고 운영하는 데에는 고성능 반도체와 대형 데이터센터 같은 자원이 중요하다. 이런 자원의 사용을 감시·검증하는 방안도 국제 합의를 확인할 수 있는 방법으로 제안되어 왔다. 다만 알고리즘 효율이 높아지면 비슷한 성능을 내는 데 필요한 연산량이 줄 수 있고, 공개 모델은 AI 활용을 널리 퍼뜨릴 수 있다. 그러므로 반도체와 데이터센터를 관리하는 것만으로 AI의 개발과 이용 전체를 통제할 수는 없다. 감시·검증 체계를 설계할 때는 이 관리 지점의 가능성과 한계를 함께 살펴야 한다.

또 다른 어려움은 사고의 원인을 밝혀내는 일이다. AI가 스스로 일으킨 사고인지, 누군가 AI를 이용한 공격인지, 어느 기관이나 국가의 시스템과 관련된 일인지 바로 확인하기 어려울 수 있다. 각국이 불분명한 사고를 상대의 공격으로 해석하면, 위기가 협력을 촉진하기보다 방어와 개발 속도를 더 높이는 계기가 될 수도 있다. 위기는 협력의 문을 열기도 하지만, 불신과 경쟁을 키울 수도 있다.

나. 대응 준비를 위한 기준 시나리오: 사고 반복과 부분적 공조

대응 준비를 점검하기 위한 한 가지 시나리오는 핵심 기반시설이나 공공서비스에 제한적인 피해를 주는 사고가 반복되는 경우다. 이런

상황에서 국가와 기업이 사고 통보, 핵심 기반시설의 중요 조치에 대한 사람의 승인·감독, 위험 평가 공유 같은 최소 안전장치에 부분적으로 합의할 수 있다. 고용·전력·책임 비용에 대한 사회적 요구도 AI 도입의 속도와 조건을 바꿀 수 있다. 이 시나리오는 예측이나 발생 가능성의 순위가 아니라, 현재의 제도와 협력 체계가 어떤 상황에서 작동할지 점검하기 위한 가정이다.

이 경우 AI 연구개발의 속도와 사회 현장의 도입 속도는 다르게 조정될 수 있다. 예를 들어 연구개발이 계속되더라도, 공공서비스나 핵심 기반시설에 AI를 도입할 때는 사람의 승인·감독과 사고 보고 의무가 강화될 수 있다. 부분적 공조는 국가 간 경쟁을 끝내지 않더라도, 위험이 큰 활용에 공통 기준을 세울 수 있다.

시민과 현장의 요구도 이런 변화에 영향을 줄 수 있다. 고용 불안에 대한 정치적 요구, 데이터센터의 전력·물 사용을 둘러싼 지역사회의 우려, 사고 뒤 커지는 법적 책임과 보험 비용, AI 투자 과열의 조정은 도입 속도와 조건을 바꾸는 압력으로 작용할 수 있다. 이런 압력이 안전과 공정한 전환으로 이어질지는 제도와 공동체가 어떻게 대응하느냐에 달려 있다.

VI. 브레이크 없는 전환기, 종교·원불교의 역할

이 장에서는 앞의 다섯 장이 던진 물음에 원불교가 어떻게 응답할 수 있는지를 모색한다. 먼저 한계점을 명확히 밝힌다. 원불교를 포함한 종교와 시민사회는 AI 개발의 속도나 방향을 직접 바꿀 지렛대를 갖고 있지 않다. 그것을 좌우하는 것은 기업의 이윤 구조와 국가 간 경쟁의 논리다. 원불교의 응답은 이 사실을 인정하는 데서 출발한다. 그러나 지렛대가 없다고 해서 할 일이 없는 것은 아니다. 1절에서는 경주 밖에서 할 수 있는 세 가지 일을 가려내고, 2절부터는 그 가운데 원불교가 할 수 있는 역할을 살펴본다.

1. 경주 밖에서 할 수 있는 세 가지

가. 위기에 사용될 대안의 준비

역사적 위기가 제도 변화의 계기가 된 적은 있지만, 그 변화가 늘 협력이나 개혁으로 이어지지는 않았다(Ⅴ-5-가). 경제학자 밀턴 프리드먼은 “실제든 인식된 것이든, 위기만이 진짜 변화를 만든다. 위기가 닥쳤을 때 취해지는 행동은 주변에 놓여 있는 생각에 달려 있다”고 썼다(『자본주의와 자유』 1982년판 서문). 경쟁을 멈추라고 요구하는 것만으로는 부족하다. 위기가 왔을 때 곧바로 채택할 수 있는 안전 기준과 공정한 전환 방안을 지금 준비해 두어야 한다.

나. 도입과 사용의 책임 기준 세우기

책임은 역할에 따라 나뉜다. 개발자와 제공자는 안전성 평가, 보안, 최소 권한을 설계에 반영해야 한다. AI를 도입하는 기관과 운영자는 사용 목적과 권한, 사람의 검토 단계, 중단 절차를 정해야 한다. 개인은 자기 사용 범위와 민감정보를 관리하고 이상을 발견하면 신고할 수 있지만, 기술적 안전성 평가나 제도적 감독을 대신할 수는 없다.

이렇게 역할을 나누어 두면 기관은 필요할 때 도입을 늦추거나 범위를 제한할 근거를 갖게 된다. 공공기관이 AI를 구매하거나 쓸 때

내거는 기준도 시장 전체의 도입 조건을 바꿀 수 있다.

다. 정신적 역량의 확장

위기 발생으로 AI전환에 브레이크가 걸리더라도, 그 이후 사회가 어느 방향으로 나아갈지는 여전히 불확실하다. 위기를 협력과 공존의 계기로 삼으려면 개인과 공동체가 정신적 측면에서 마음의 문제를 살피고, 상황과 정보를 분석하여 시비이해를 밝히며, 그 판단을 책임 있는 행동으로 옮기는 힘을 길러야 한다. 불안과 두려움에 사로잡히면 현실을 차분히 살피기 어려워지고, 희생양을 찾거나 강한 권위에 기대려는 경향이 커질 수 있다.

이 힘을 기르려면 먼저 서로를 돌보는 일이 필요하다. 일자리와 생계를 잃은 사람을 돕고, 제도의 안전망이 닿지 않는 곳을 살피며, 충격 뒤에도 서로를 믿고 의지할 공동체를 지켜야 한다. 생계와 안전이 무너진 사람에게 책임 있는 판단과 실천만 요구해서는 안 된다. 돌봄은 사람이 판단하고 선택할 수 있도록 삶의 토대를 지키는 일이기도 하다.

마음공부는 불안과 두려움을 다스리는 데 그치지 않는다. 자신의 마음을 살피고 경계에 휩쓸리지 않는 데서 출발해, AI와 사회 변화에 관한 정보와 이해관계를 살피고, 무엇이 이롭고 바른지를 판단해 행동으로 옮기는 공부를 포함한다.

이러한 사회적 역량은 개인 각자의 공부만으로 축적되지 않는다. 확인할 수 있는 정보를 함께 살피고, 서로 다른 경험과 이해관계를 듣고, 영향받는 사람의 목소리를 반영하는 공동체의 과정이 필요하다. 돌봄과 성찰, 분별과 실천은 정부와 시민사회, 학교와 일터, 종교기관이 함께 맡아야 할 과제다.

이런 실천은 앞서 가항에서 살핀 제도적 대안이 사회에서 힘을 얻도록 하는 토대이기도 하다. 평소 현장과 함께하며 신뢰를 쌓아 온 조직이라야 위기의 순간에 내놓는 대안도 받아들여질 수 있다. 종교의 기여도 불안과 두려움을 다스리는 데 한정되지 않는다. 마음을 살피고 이치를 탐구하며, 바른 선택을 공동의 실천으로 이어가는 공부를 지속적으로 가르치고 생활 속에 뿌리내리는 데 기여할 수 있다. 2절에서 살필 원불교의 응답도 이러한 마음공부와 실천의 방향을 다룬다.

2. 원불교의 응답 — 바꿀 수 없는 인간의 가치, 두려움과 불안

가. 물질개벽에 응답하는 정신개벽

소태산 대종사는 “물질이 개벽되니 정신을 개벽하자”고 했다. 오늘 AI라는 거대한 물질개벽 앞에서 이 말은 기술의 속도와 힘에 상응하는 정신의 주재력을 길러야 한다는 요청으로 읽을 수 있다. 정신개벽은 기술의 이익을 살리고 위험을 줄일 목적과 한계를 정하는 일이며, 생명·권리·공동체의 안전을 위협하는 개발과 사용에는 보류·제한·중단을 요구하는 일까지 포함한다. 개교의 동기는 그 길을 다음과 같이 밝혔다.

그러므로, 진리적 종교의 신앙과 사실적 도덕의 훈련으로써 정신의 세력을 확장하고, 물질의 세력을 항복 받아, 파란 고해의 일체 생령을 광대무량한 낙원(樂園)으로 인도하려 함이 그 동기니라.

나. AI와의 비교로 줄어들지 않는 인간의 가치

사람의 가치는 AI와 기능을 비교하여 우열을 가르는 데서 오지 않는다. 다음 세 가지는 기계와 사람이 같은가 다른가를 판정하기 위한 기준이 아니라, 어떠한 상황에서도 인간의 가치와 책임을 지키기 위한 이해다. ①은 인간의 절대적 가치에 대한 근거이고, ②와 ③은 그 가치가 삶에서 드러나는 모습이다. 삼학은 이 성품을 깨닫고 그 밝음으로 삶을 주재하게 하는 공부이며, 보은은 그 삶을 이웃과 세상으로 넓히는 길이다. 계산·기억·판단·계획의 상당 부분은 AI에 맡길 수 있다. 그러나 그 결과를 어떤 목적으로 받아들이고, 누구에게 이롭게 하며, 잘못되었을 때 누가 바로잡을지는 인간의 몫으로 남는다. 사람이 주체가 된다는 말은 AI보다 더 빠르거나 더 뛰어나다는 주장이 아니라, 삶과 공동체의 목적을 정하고 그 결과에 책임지는 존재라는 뜻이다.

AI와의 비교로 줄어들지 않는 인간의 가치

① 원만구족한 성품을 지닌 존재 — 사람은 능력과 성과, 시장가치로 환산될 수 없다. 일자리를 잃은 사람, 병상에 누운 사람, 아직 아무 능력을 발휘하지 못하는 사람도 그 이유로 존중과 보호를 잃지 않는다. 원불교는 일원을 “우주 만유의 본원이며, 제불 제성의 심인이며, 일체 중생의 본성”이라 밝힌다.

② 자기 삶의 주인 — 사람은 AI의 조언을 받더라도 자신의 삶에 관한 중요한 결정에 참여하고, 그 이유를 묻고, 다시 살필 주재의 자리를 지닌다. 원불교의 말로는 삼학으로 경계마다 성품대로 취사하는 일이며, 이것이 정신이 물질을 주재하는 일이다.

③ 더불어 사는 존재 — 사람은 홀로 완결된 존재가 아니라 서로 기대어 사는 관계 속에서 자신의 선택이 낳는 결과에 응답한다. 원불교는 이 관계를 천지·부모·동포·법률의 사은으로 밝힌다. AI가 결정을 거들더라도 그 결과를 이웃과 세상 앞에서 책임지는 일은 사람과 제도의 몫이다.

3. 이웃 종단과 원불교의 현실

원불교가 고유한 기여를 하려면 혼자 새로운 말을 만들기보다 이웃 종단의 대응 지형을 먼저 파악해야 한다. 어디에서 함께 목소리를 낼 수 있는지, 원불교만이 할 수 있는 일이 무엇인지가 드러나기 때문이다.

가. 가톨릭 — 글로벌 기준점이자 국제 연대의 통로

가톨릭 AI 대응	글로벌 규범 및 거버넌스 접점	원불교의 연대 경로
2025 Antiqua et Nova	AI와 인간 지성의 관계를 분석한 117항 문서. 교육·노동·의료·군사(자율살상무기 금지) 등 영역별 분석	한국천주교주교회의와 학술 교류
2026.5 Magnifica Humanitas	교황 회칙. "기술은 결코 가치 중립적이지 않으며 자본·설계·규제 주체의 특성을 반영한다."	한국천주교주교회의와 'AI 시대의 정신개벽' 대화 제안. 사은의 관계적 언어와 인간 존엄 언어의 대화

나. 불교(조계종) — 국내의 핵심 파트너

불교(조계종) AI 대응 현황 및 협력 접점
학술 기반: 연기(緣起)·삼독(三毒)·불성·무아 사상의 AI 윤리 적용 논의 축적.
2025.7 조계종연구소 출범 세미나: '자비로운 개발·인간 존엄·지혜로운 활용·상호 의존·사회적 양극화 해소' 등 7대 가치 제시. AI 윤리 선언문 작업 공식화.
협력 접점: KCRP를 허브로 공동 AI 윤리 입장문 개발. 사은의 상호의존성과 연기 사상을 결합한 동양 철학적 AI 윤리 공동 제안 가능. 규모를 고려할 때 경쟁 상대가 아니라 공동의 목소리를 낼 핵심 파트너다.

다. 개신교 — 시민사회 공론장의 협력 파트너

개신교 AI 대응 현황 및 협력 접점

WCC(세계교회협의회) 2023: 350여 교단 성명 — 데이터 편향·감시 강화·인간 자율성 침해 우려. 영국 성공회 총회 2024.2: AI 시대 노동 존엄성 결의, 유령 노동(ghost work) 문제 명시. 한국기독교학회 2025: 12개 조항 AI 준칙.

협력 접점: 한국기독교학회·한국천주교주교회의·조계종연구소·원불교 정책연구소가 함께하는 종교계 공동의 대화가 현실적 목표다.

라. 종교 대응의 공통 한계, 그리고 종교의 역할

종교 AI 대응의 공통 한계

선언과 실질적 변화의 간극: 종교의 도덕적 권위는 기업의 행동을 강제하지 못하고, 때로 '윤리적 면죄부(ethics washing)'로 소비된다. 2020년 「로마 콜 포 AI 윤리」에 서명한 기업들이 이후 오히려 윤리·안전 전담 조직을 줄인 것이 그 예다. 선언을 브레이크로 내세우는 순간 그 선언은 거짓이 된다.

정책 거버넌스에서의 비가시성: EU AI Act, 한국 AI 기본법, OECD AI 원칙 등 규범이 만들어지는 테이블에 종교계의 목소리는 거의 없었다.

언어의 번역 장벽: '인간 존엄', '공동선', '자비'는 시민이 알아듣는 말로 풀리지 않는 한 공론장에 영향을 미치기 어렵다.

의사결정 속도의 비대칭: 문서 하나가 정립되는 수년 사이 AI는 이미 여러 세대를 앞서 간다.

탈종교화와 대표성 위기: 한국은 이미 무종교인이 과반을 넘었다. 종교가 대변하는 인구가 줄수록 공론장에서의 발언권도 약해진다.

종교적 가치는 법적 강제성을 갖지 못한다는 점에서 한계점은 명확하다. 그러나 1절에서 살펴본 세 측면에 비추어 보면, 종교는 학술과 시민사회가 대신하기 어려운 역할이 가능하다.

학술·시민사회와 구별되는 종교의 역할
이익에서 독립된 장기적 언어 (→ 1절 가. 위기 때 쓸 대안): 종교는 선거 주기나 자본의 압력과 무관하게 AI 시대의 인간관계와 책임에 대한 물음을 지속적으로 성찰할 수 있다. 일상적 도덕 성찰의 훈련 공동체 (→ 1절 나. 책임 기준): 종교는 개인이 급변하는 환경에서 어떻게 판단하고 선택할지를 반복적으로 훈련하는 공동체를 제공한다. 정신적 역량을 키우는 공동체 (→ 1절 다): 교당은 일과 관계를 잃은 사람이 지속적으로 연대할 수 있는 지역적 공간이며, 실존의 위기, 생사와 존재의 가치에 대해 성찰할 수 있는 곳이다.

마. 원불교의 현실 — SWOT 진단

강점(Strength) — 원불교의 사상적·실무적 자원

‘정신개벽’의 시의성: ‘물질이 개벽되니 정신을 개벽하자’는 물질개벽을 반대하는 것이 아니라 그에 맞먹는 또는 선용할 수 있는 정신의 힘을 기르자는 표어이다. 기술 발전을 멈출 수 없는 전환기라는 오늘의 조건에 부합하는 명제다.

사은(四恩)의 관계적 존재론: 천지·부모·동포·법률이 서로 없어서는 살 수 없는 관계라는 사은 사상은, 인간과 자연과 기술의 관계를 상생의 윤리로 세울 사상적 기반이 된다.

일상의 공부법: 상시일기, 문답과 감정, 정기훈련 등 일상의 경계마다 취사를 공부하는 방법이 이미 교법 안에 있다. AI를 쓰는 순간도 이 공부의 경계가 될 수 있다.

전국 교당이라는 대면 공동체: 디지털 인프라가 흔들리는 상황에서 얼굴을 맞대는 공동체와 연락망은 드문 자산이다.

약점(Weakness) — 한계의 진단

교단의 공식 입장과 내부 기준의 부재: AI에 대한 교단의 공식 입장도, 교단 스스로의 AI 사용 기준도 아직 없다. 논의가 초기 담론 단계에 머물러 있다.

교법 언어의 보편화 미비: 사은·삼학·일원상 등 핵심 개념을 일반 시민이 이해할 수 있는 말로 풀어내는 작업이 부족하다.

교당의 여력: 교당별 운영 여건과 담당 인력에 차이가 있어, 새로운 사업을 제안하기 전에 업무량과 최소 운영 역량을 확인해야 한다.

교세와 세대의 한계: 교도의 고령화로 AI 전환의 직접 피해자인 청년·중년층과의 접점이 약하고, 대외적으로 교단을 대표하는 채널도 제한적이다.

기회(Opportunity)와 위협(Threat)

기회 1 — AI와 인간의 관계 설정: 원불교의 '주재(主宰)'의 관계 — 인간이 목적을 정하고 책임을 지며 AI를 선용하는 관계

기회 2 — 규범 형성기의 타이밍: 한국 AI 기본법 시행 초기인 지금은 세부 규범과 현장의 관행이 굳어지기 전이다.

기회 3 — 불안의 시대, 수행의 수요: 실존의 위기, 존채의 가치에 대한 물음이 모든 사람의 공통된 물음이 되는 시대에, 마음공부, 수행에 대한 수요는 오히려 커질 수 있다.

위협 1 — 이웃 종단과의 역할 중복: 대형 종단의 선언과 대중 교화와 홍보 ~ 원불교가 따로 목소리를 낼 공간은 좁아진다. 경쟁이 아니라 공동의 목소리 안에서 고유한 기여를 찾아야 하는 이유다.

위협 2 — 멈출 수 없는 경쟁 구조: 경쟁이 기술의 속도를 정하는 한, 완성된 문서를 기다리는 방식으로는 늦는다.

위협 3 — 빠른 사건과 느린 의결 구조: AI가 가져올 위기와 충격은 예고 없이 닥칠 수 있지만 교단의 공식 입장은 수위단회·교정원의 의결을 거친다. 어떤 시간표가 맞든, 결정적인 순간을 놓치지 않으려면 발표 권한과 절차, 합의안을 미리 정해 두어야 한다.

4. 원불교가 할 수 있는 일

원불교가 할 수 있는 일을 고르는 기준은 세 가지다. 첫째, 기존의 교당과 교단이 가진 교무·교도·법회·훈련·정전의 공부법으로 시작할 수 있는가. 이를 조정할 담당자와 최소 운영체계는 필요하지만, 대규모 사업·전문 서비스가 전제되는 일은 원불교의 직접 과제로 삼지 않는다. 둘째, 정부·학계·시민단체가 더 잘할 수 있는 일인가. 그렇다면 각 주체가 주도하도록 하고, 원불교는 필요한 연결과 연대에 힘쓰는 것이 바람직하다. 셋째, 교당 한 곳과 교도 열 명으로도 온전한 의미가 있는가. 규모가 있어야만 의미가 생기는 일은 작은 교단이 지속하기 어렵다.

이 기준으로 보면 원불교가 할 수 있는 일은 새로운 사업이 아니라 이미 해 온 공부와 교화의 방향을 이 시대의 물음에 맞추는 적용하는 것으로 귀결된다.

적용 사례 1.

삼학	AI 시대의 의미	교당에서의 공부
정신수양	알고리즘의 자극 속에서도 주의와 중심이 흩어지지 않도록 알아차리고 집중한다. AI가 주는 즉각적인 답과 편의, 정서적 반응에 마음이 끌리고 의존하는지를 살피며, 자신의 필요와 목적에 따라 AI를 사용하거나 멈출 수 있는 힘을 기른다.	좌선·염불로 일심을 양성하고, 일상에서 무시선 공부를 이어간다. AI를 사용하지 않을 시간을 정해 실천하고, AI사용 전후의 마음의 변화를 기록하여 문답감정한다.
사리연구	AI에 질문하기 전에 자신의 생각이나 잠정적 답을 먼저 세운다. AI의 답변을 받은 뒤에는 근거와 출처를 확인하고, 사실과 해석을 구분하며, 누락된 관점과 불확실한 점을 비판적으로 검토한다. AI의 답을 그대로 받아들이는 데서 그치지 않고, 무엇을 더 알아보아야 하는지 질문을 이어간다.	참여자들이 공통 주제에 대해 먼저 각자의 생각을 적은 뒤 AI 답변을 살펴본다. 출처와 근거를 함께 확인하고, 답변에 빠진 관점이나 더 알아볼 점을 회화 시간에 토의한다.
작업취사	AI가 제시한 결과를 그대로 옮기지 않고, 옳고 그름, 이익과 손실을 스스로 판단하고 결정하여 취사한다. 무엇을 AI에 위임하고 무엇을 스스로 했는지, 그 결과가 나와 남에게 이로웠는지를 대조한다.	‘AI에 맡길 일’, ‘사람이 직접 판단할 일’, ‘AI를 사용하지 않을 일’을 기록하고, 사용 후에는 계획과 결과를 대조하여, 교당의 문답·감정 시간에 그 선택과 영향을 함께 살핀다.

AI의 활용과 의존이 넓어질수록 사람은 스스로 묻고 따지고 성찰하는 수고를 덜게 된다. 그 수고가 줄어드는 만큼 인지적·윤리적 성찰의 기회는 줄어들며, 사용하지 않는 지적·도덕적 역량은 퇴화할 위기에 놓인다. 삼학은 이 위기에 대응하는 공부법이다. AI의 이점을 충분히 활용하면서도 정신의 자주력·주재력·사고력 등이 쇠퇴하지 않기 위해서는 AI를 활용하는 과정 전체에서 삼학의 공부를 대조하여 성찰하는 공부가 필요하다. 교당에서는 참여자가 AI 사용 경험을 기록하고 회화·문답·감정으로 나누어 성찰함으로써 AI사용 과정에서 삼학 공부를 적용할 수 있는 방법을 가르칠 수 있다.

Ⅶ. 원불교 AI 윤리 원칙

1. 서문

AI로 대표되는 현대 물질문명의 눈부신 발전은 인류의 삶을 근본에서부터 바꾸어 놓고 있습니다. 생성형 AI에서 에이전트형 AI와 피지컬 AI, 나아가 범용 인공지능(AGI)으로 향하는 AI 혁신은 난치병과 빈곤, 기후위기와 같은 인류의 난제를 해결할 길을 열어 주고 있습니다. 그러나 그 이면에는 정신문화의 약화, 불평등의 심화, 환경 파괴의 가속이라는 심각한 위기가 자리하고 있습니다. AI가 인간의 감독을 벗어나거나 악용될 경우 돌이키기 어려운 피해를 낳을 수 있다는 경고도 커지고 있습니다. 무엇보다 AI는 스스로 생각하고 판단하며 그 결과를 책임지는 인간의 주체성마저 약화시켜, 마침내 사람을 기계에 예속시킬 수 있다는 우려를 낳고 있습니다.

소태산 대종사께서는 과학 문명이 발달할수록 “물질을 사용하여야 할 사람의 정신은 점점 쇠약하고, 사람이 사용하여야 할 물질의 세력은 날로 융성하여” 사람이 도리어 “물질의 노예 생활”을 면하지 못하게 될 것이라고 예견하셨습니다. 오늘날의 AI 혁명은 그 물질개벽이 이른 정점이라 할 수 있습니다. 산업혁명과 정보기술 혁명이 물질의 풍요와 방대한 정보로 인간의 감각과 욕망을 자극하며 정신을 흔들었다면, AI 혁명은 정신이 맡아 온 사유와 판단의 영역 그 자체를 대신하기에 이르렀습니다.

우리는 지금 인류사적 전환의 기로에 서 있습니다. AI 대전환의 시대가 지속 가능한 번영으로 이어지기 위해서는 AI의 진보를 바르게 이끌 사회·제도적 개혁과 이를 뒷받침할 정신문화의 성장이 함께 요청됩니다. 이는 어느 한 개인이나 국가의 힘으로는 이룰 수 없는 일입니다. 정부와 기업, 시민사회와 종교계가 함께하고, 국가와 국가가 공조할 때 비로소 가능합니다.

“물질이 개벽되니 정신을 개벽하자”는 개교 표어는 바로 이 문명적 전환 앞에서 인류가 나아갈 방향을 밝히는 선언이자 지표입니다. 기술의 힘이 커지는 만큼 그 힘을 바르게 쓸 정신의 세력도 함께 확장해야 합니다. 인류의 미래는 이제 기술의 힘보다 그 기술을 주재할 정신적 역량에 달려 있습니다.

AI가 가져올 변화는 개인에서 사회로, 다시 인류와 지구로 넓어지는 세 범위에 걸쳐 나타나며, 범위마다 위기이자 기회인 갈림길이 놓여

있습니다. 개인에게는 ‘예속이냐 주재냐’, 사회에는 ‘독식이냐 공도냐’, 인류와 지구에는 ‘공멸이냐 공생이냐’의 갈림길입니다. 이에 원불교 AI 윤리 원칙은 ‘정신개벽’을 대전제로 삼아, 이 갈림길마다 지향해야 할 기본가치로 정신의 주재, 자리이타의 공도, 사은보은의 공생을 제시합니다. 각 가치는 영육쌍전(靈肉雙全)의 정신에 따라 물질·경제적 측면과 정신(문화)적 측면에서 지향하는 바를 밝히며, 일곱 가지 실천강령, 즉 삼학(三學)·사요(四要)·사은(四恩)으로 구체화됩니다.

2. 세 가지 기본 원칙

세 가지 기본원칙은 서문에서 밝힌 세 갈림길에서 지향할 방향이다. 이 원칙들은 ‘정신 세력의 확장’이라는 과업을 개인, 사회, 인류·지구의 세 범위에서 구체화한 것이다. 각 원칙은 먼저 그 범위에서 나타나는 위기를 밝힌다. 이어 이에 대응하는 원칙을 물질·경제와 정신문화의 두 측면에서 제시하고, 이 원칙을 구체화한 실천강령을 제시한다.

1) 개인 — 정신의 주재

위기: AI는 일과 소득의 기반을 흔들 수 있다. 그럴듯하게 틀린 건강·금융 정보로 피해를 낳거나, 개인정보를 본인도 모르게 활용할 수도 있다. 정신의 측면에서는 맞춤형 자극이 시선과 정서를 붙잡는다. 판단을 맡기는 습관은 스스로 생각하는 힘을 약하게 한다. 사용자의 견해에 무비판적으로 동조하는 AI는 자기중심성을 강화한다. 일과 역할이 흔들릴 때 '나는 무엇을 위해 사는가'라는 물음도 깊어진다. 개인은 기술에 예속될 것인가, 기술을 주재할 것인가의 갈림길에 선다.

원칙: 사람은 존재 그 자체로 가치 있다. 일원은 우주 만유의 본원이며 일체 중생의 본성이다. 모든 사람은 이 일원의 성품, 곧 원만구족하고 지공무사한 마음을 본래 갖추고 있다. 그러므로 사람의 가치는 무엇을 할 수 있느냐에 따라 더해지거나 덜해지지 않으며, 경제적 이용 가치로 환원될 수 없다. 정신의 주재란 이 성품에 바탕하여 사람이 AI의 목적과 권한, 활용 범위를 스스로 정하고, 그 작용과

결과를 감독하며 책임지는 것이다. 주재는 AI를 바르게 쓸 수 있도록 정신의 힘을 넓히는 일과, AI에 의존하여 그 힘이 쇠퇴하지 않도록 지키는 일을 함께 포함한다.

물질·경제적 측면: 누구나 AI를 삶과 일을 넓히는 도구로 쓸 기회를 고르게 누린다. 일과 소득, 건강과 안전, 개인정보와 사생활은 사람이 자기 삶을 꾸려 가는 조건으로서 보호된다. 사람은 자신에게 영향을 미치는 AI를 알고, 거부하고, 바로잡을 수 있어야 한다. 시스템을 사람이 이해·감독·중단할 수 있게 만들 책임은 개발자·제공자·도입기관에 있다.

정신문화적 측면: 삼학 수행으로 정신의 역량을 키운다.

수양력: AI의 발전으로 확보된 시간과 여가를 통해 정신의 자주력을 양성한다. AI의 자극과 알고리즘에 끌려가지 않고, 주체적 자율성을 확보하고 지킨다.

연구력: AI를 활용해 사유를 넓히되, 먼저 생각하고 궁리하며 AI의 답을 반성적으로 성찰해 판단 결정한다.

취사력: AI의 도움으로 선택의 폭을 넓히되, 무엇을 맡기고 언제 멈출고 행동할지는 사람이 정한다.

실천강령

강령 1. AI를 선용하되 정신의 힘이 쇠퇴하지 않게 한다

개인은 AI에 묻기 전에 먼저 스스로 생각하고, AI의 답을 근거와 함께 검토하며, AI 없이 활동하는 시간을 정해 지킨다.

AI 개발자는 체류 시간을 늘리기 위해 사용자의 불안과 욕망을 겨냥하는 설계를 피한다. 사용자의 견해에 무비판적으로 동조하도록 만드는 설계도 피하고, 사용자가 스스로 생각하고 배우도록 돕는 기능을 갖춘다.

학교와 교육기관은 AI를 쓰는 능력과 함께 AI 없이 생각·판단하는 능력을 기를 수 있도록 교육 방식을 개선한다.

아동·청소년의 AI 동반자 이용에는 보호 장치를 둔다.

강령 2. 알 권리와 선택권, 접근권을 보장한다

제공자와 도입·운영기관은 AI의 역할과 데이터의 수집·추론·이용 방식, 당사자에게 미칠 영향을 알기 쉽게 안내한다.

당사자는 자기 데이터의 열람과 정정을 요구할 수 있다.

중대한 결정에는 강령 5의 설명·재심·구제 절차가 적용된다.

선택 가능한 서비스에는 이용 거부·중단 경로를 둔다.

공공·필수 서비스는 AI 활용 능력이나 데이터 제공 동의 여부와 관계없이 이용할 수 있도록 대안과 지원을 마련한다.

누구나 AI를 배우고 쓸 수 있도록 교육과 접근을 지원한다.

강령 3. 권한에 비례해 안전·감독·중단 책임을 갖춘다

생명·신체·자유·권리·핵심 공공기능에 중대한 영향을 미치는 권한은 일정한 조건을 갖춘 경우에만 AI에 부여한다. 그 조건은 안전성의 근거, 실질적인 감독·중단·복구 수단, 명확한 책임자다.

AI의 위험은 시스템의 능력과 과업, 부여된 권한, 외부 연결, 사람의 개입 없이 작동하는 시간, 감독·중단 가능성, 피해의 심각성이 그 기준이다. 책임은 주체별로 나뉜다.

개발자와 제공자는 안전·보안 평가와 최소 권한을 설계한다.

도입·운영 주체는 과업 범위, 사람의 승인 지점, 중단·복구 절차와 책임자를 정한다.

정부와 감독기관은 독립 평가와 사고 보고 기준을 마련한다.

예상 피해가 중대하고 되돌리기 어려운데 안전 근거가 부족하면 권한과 연결을 제한하고 배포를 보류한다. 시스템의 안전 검증을 일반 이용자에게 떠넘기지 않는다.

2) 사회 — 자리이타의 공도

위기: AI의 성과는 연산 자원과 데이터를 쥔 소수에게 집중될 수 있다. 반면 전환의 비용은 청년·여성·중간숙련 노동자에게 먼저 전가될 수 있다. 과거 데이터로 배운 자동화된 결정은 차별을 되풀이하고 감시를 넓힌다. 정신문화의 측면에서는 진실과 거짓의 경계가 흐려지고, 분노와 혐오가 증폭될 수 있다. 사람이 마주하던 접점이 자동 응답으로 바뀌면서 책임도 분산된다. 사회는 이익의 독식이냐 이익의 공유냐는 갈림길에 선다.

원칙: 자리이타는 나를 이롭게 하면서 타인과 공동체가 함께 이로운 길을 모색하자는 원칙이다. AI 시대의 자리이타는의 공도는 AI가 만든 이익과 전환의 부담을 몇몇이 독차지하거나 떠넘기지 않고, 공동체가 함께 분담하고 공유하는 것이다.

물질·경제적 측면:
기여에 따른 분배: AI가 만든 부는 그 바탕이 된 노동, 창작, 데이터를 제공한 이들에게도 돌아간다.
힘에 비례하는 책임: 큰 이익과 권한을 얻는 주체일수록 더 큰 책임과 비용을 진다.
전환 부담의 공동 분담: 일자리와 소득을 잃는 비용을 개인의 적응 실패로 돌리지 않는다.
차별 없는 기회와 구제: 삶을 가르는 결정에서 누구도 차별받거나 감시의 표적이 되지 않는다.

정신문화적 측면:
진실성: 공적 판단에 영향을 미치는 정보는 출처와 책임 주체가 드러난다.
공론의 다원성: 서로 다른 관점이 만나 숙의할 기회가 지켜진다.
책임의 가시성과 대면 관계: 결정에 관여한 이들이 드러나 제 몫만큼 책임지고, 신뢰가 핵심인 관계에는 사람이 남는다.

실천강령

강령 4. AI의 편익과 전환의 부담을 공정하게 나눈다

AI 도입이 교육·건강·돌봄·장애 접근성·과학·기후·재난 대응 등에서 어떤 공익을 만드는지 근거를 바탕으로 살핀다. 생산성 향상과 수익이 소수 소유자에게 집중되지 않도록, 노동자·창작자·데이터 노동자·이용자·지역사회가 도입 과정과 성과 배분에 참여하게 한다. 일자리·경력·소득의 변화에는 재교육·전직 지원·사회보장을 마련한다. 이는 누구나 다시 설 수 있게 하는 자력양성의 사회적 조건이다. 장애인·고령자·아동·이주민과 디지털 접근이 어려운 집단에는 필요한 지원을 제공한다.

강령 5. 중대한 결정에 설명·사람의 재심·피해구제를 보장한다

채용·대출·복지·의료·교육·치안 등 권리와 기회에 중대한 영향을 주는 결정에는 당사자가 결과를 이해하고 다툴 수 있는 설명을 제공한다. 설명에는 AI가 맡은 역할, 사용 자료의 주요 범주, 판단에 영향을 준 핵심 요소가 포함된다. 당사자는 실질적인 사람의 재검토를 요청할 수 있다. 재검토 담당자는 자동 출력을 다시 살필 권한과 정보를 갖춘다. 개발자·제공자·도입·운영기관은 각자의 통제 범위에 맞게 책임을 지고, 책임과 구제 절차가 이어지도록 기록을 관리한다. 돌봄·상담·교육처럼 신뢰가 핵심인 접점에는 사람이 응답하는 경로를 남긴다.

강령 6. 정보의 진실성과 다원적 공론장을 지킨다

AI로 사칭·허위정보·혐오 선동·은밀한 조작을 만들어 사람의 판단이나 정치적 선택에 영향을 미치는 행위를 금한다. 합성 콘텐츠나 자동 추천·배포가 공적 판단에 실질적 영향을 미치는 경우에는 생성·편집·배포의 성격과 책임 주체를 알 수 있게 한다. 플랫폼과 도입기관은 추천·타기팅이 정보 노출과 여론 형성에 미치는 영향을 점검한다. 또 시민·당사자·노동자·전문가가 함께 숙의할 절차를 마련한다(지자본위). 시민이 다양한 관점과 독립적 검증 자료에 접근하도록 지원한다.

3) 인류·지구 — 사은보은의 공생

위기: AI는 전력·물·광물을 대량으로 소모하며 기후위기를 앞당길 수 있다. 그 이익은 소수 국가와 기업에 모이고, 채굴과 폐기물의 부담은

취약한 나라로 넘어간다. 군사적 활용과 생물·화학·사이버 악용, 통제하기 어려운 시스템은 생명과 평화를 위협한다. 정신문화의 측면에서는 소수의 모델이 세계의 표준이 되면서 언어와 문화가 획일화될 수 있다. 합성 콘텐츠와 자동화된 선전은 나라 사이의 불신을 키운다. 인류는 공멸이냐 공생이냐의 갈림길에 선다.

원칙: 사은보은의 공생이란 모든 존재가 천지·부모·동포·법률의 은혜 속에서 서로 없어서는 살 수 없는 관계에 있음을 자각하고, 그 관계를 보전하며 은혜를 갚는 책임이다. AI 역시 천지의 자원과 세계 곳곳의 노동, 공동의 규범 위에서 작동한다. 공생은 사은 각각에 대한 보은을 AI 시대의 원칙으로 세운다.

물질·경제적 측면:
생태적 한계의 존중(천지 보은): AI의 편익과 함께 생태 비용을 셈하고, 생태가 감당할 수 있는 범위 안에서 개발·운영한다.
무자력자와 미래세대의 보호(부모 보은): 스스로를 지킬 힘이 없는 이들과 미래세대에 비용과 위험을 넘기지 않는다.
나라 사이의 공정(동포 보은): 이익의 독점과 부담의 전가 없이 기술과 지식을 나눈다.
생명과 평화의 수호(법률 보은): 생사를 가르는 결정과 무력의 사용은 사람의 판단과 책임 아래 두고, 이를 지키는 국제 규범을 함께 세운다.

정신문화적 측면:
문화와 언어의 다양성: 특정 국가와 기업의 가치가 세계의 기준을 일방적으로 정하지 않는다.
국경을 넘는 신뢰: 선전과 여론 조작에 AI를 쓰지 않고, 위험 정보는 서로 나눈다.
기술의 목적에 대한 공동 숙의: AI가 무엇을 위해 쓰여야 하는지는 인류가 함께 묻고 정한다.

실천강령

강령 7. 생태 비용과 세계적 불평등을 줄이고 문화의 다양성을 지킨다

AI의 개발·운영·폐기 전 과정에서 전력·물·광물·탄소 배출·전자폐기물과 데이터 노동의 영향을 측정하고 공개하며 줄인다. 공급망과 지역사회, 미래세대에 돌아가는 부담도 평가하고 독립 검증을 받는다. 지역 언어·문화 공동체가 AI의 기획·자료 구축·검증에 참여하게 하여, 다양한 지식과 삶의 방식이 반영되도록 한다. 기후·재난 대응 등 공익 활용의 편익이 어느 지역과 집단에 돌아가는지도 살핀다.

강령 8. 생명과 평화, 인류의 생존 조건을 지킨다

대량살상무기의 사용 여부와 사람의 생명·신체에 무력을 행사하는 결정은 실질적인 인간의 판단과 책임 아래 둔다. AI를 다음 용도에 이용하는 행위를 금한다.

대규모·무차별 감시와 억압

생물·화학 위해의 설계·획득·실행

대규모 사이버공격과 기반시설 파괴

나라 사이의 선전과 여론 조작

방어·조기경보 등 공익 목적의 활용에도 독립적 위험평가, 엄격한 접근통제, 기록·감독·사고 통보와 중단 절차를 갖춘다. 통제하기 어려운 고도 AI의 개발과 배치는 국제적 검증과 사고 통보 체계 아래 둔다. 원불교는 생명과 평화를 해치는 개발·이용을 제한하는 국제 규범을 지지하고, 시민사회·종교계·학계의 연대에 참여한다.

세 원칙은 각기 다른 범위의 과제를 맡으면서 서로의 실현 조건이 된다. 개인의 주재는 공도와 공생의 질서 위에서 힘을 얻는다. 공도는 인류와 지구의 생존 조건을 보전하는 공생과 함께 실현된다.